

Modeling Symbolic Music with Natural Language Processing Approaches

Analyzing and Experimenting with Multiple Levels of Representations

Dinh-Viet-Toan LE

PhD Thesis Defense

November 3, 2025

Xavier HINAUT

Cheng-Zhi Anna HUANG

Emmanouil BENETOS

Chloé BRAUD

Marius BILASCO

Patrick BAS

Louis BIGO

Marc TOMMASI

Mikaela KELLER

Chargé de recherche Inria, Inria Bordeaux

Associate Professor, MIT

Reader & Director of research, QMUL

Chargée de recherche CNRS, IRIT

Professeur des universités, CRISTAL

Directeur de recherche CNRS, CRISTAL

Professeur des universités, LaBRI

Professeur des universités, CRISTAL

Maîtresse de conférence, CRISTAL

Reviewer

Reviewer

Examiner

Examiner

Examiner

Jury president

Co-director

Co-director

Co-supervisor

323

Fl.

Cl.

Fg.

Vln.1

Vln.2

Vla.

Vc.

Cb.

beethoven5

L.v. Beethoven, *Symphony No. 5*, Mvnt. 1. Bars 323–330.

“A musical phrase is passed through the instruments like a **dialog**.”

323

Fl.

Cl.

Fg.

Vln.1

Vln.2

Vla.

Vc.

Cb.

beethoven5

L.v. Beethoven, *Symphony No. 5*, Mvnt. 1. Bars 323–330.

- “Call-and-response”
- “Musical discourse”
- “A composer’s vocabulary”
- ...
- “Musical language”

Jackendoff, *Parallels and nonparallels between language and music*, 2009.

Music & language: some fundamental similarities...

Modalities

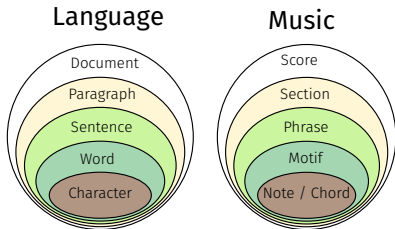
	Auditory	Written
Language	Speech	Text
Music	Audio	Score

Sequential representation

<Model_><ing> <music_><al> <content> <with> <NLP>



Segmentation



“Grammar”

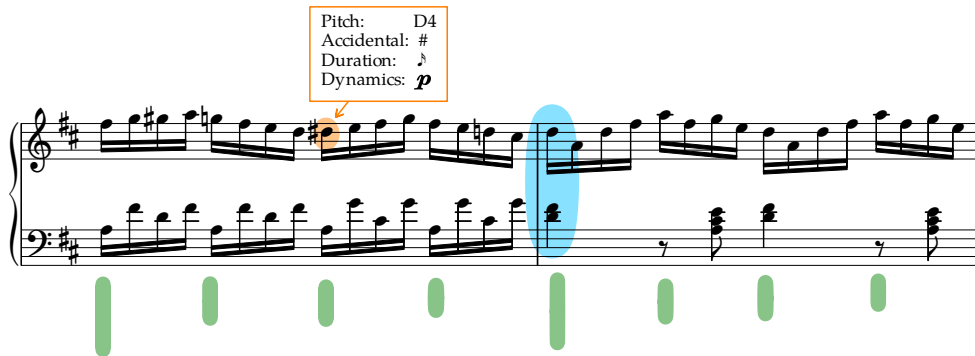
- Grammatical trees
- Tonal theory (GTMM [Lerdahl & Jackendoff 1987])

Expectancy

- The cat is eating the [X]
- Functional harmony

Music & language: ...and major differences

- Temporal dimension in language & music
- Polyphony in music
- Polymorphism of musical symbols

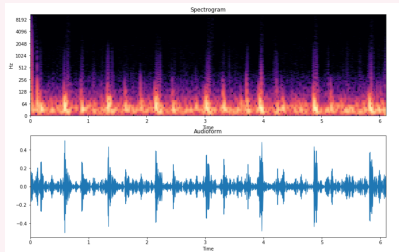


A musical score snippet in treble and bass clef, key of D major. The treble staff contains a melodic line with eighth and sixteenth notes. The bass staff contains a bass line with eighth and sixteenth notes. A callout box points to a specific note in the treble staff. Below the staff, there are green vertical bars indicating the temporal dimension of the music.

Pitch: D4
Accidental: #
Duration: ♪
Dynamics: *p*

Symbolic music

Audio



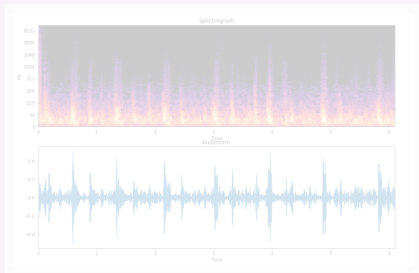
- Waveforms, spectrograms...
- Timbre, performance expressivity...

Symbolic music

- Score (sheet music):
 - Notes, chords, instruments, dynamics, articulations...

Symbolic music

Audio



- Waveforms, spectrograms...
- Timbre, performance expressivity...

Symbolic music

A snippet of a musical score for a symphony orchestra. The score includes staves for Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), Violoncello (Vc.), and Contrabass (Cb.). The tempo is marked as '♩ = 80'. The key signature has one sharp (F#). The score shows various musical notations including notes, rests, and dynamic markings such as 'p' (piano) and 'pizz.' (pizzicato). The Viola part has 'arco.' (arco) markings. The Vc. and Cb. parts have 'pizz.' markings.

- Score (sheet music):
 - Notes, chords, instruments, dynamics, articulations...

Natural Language Processing (NLP)

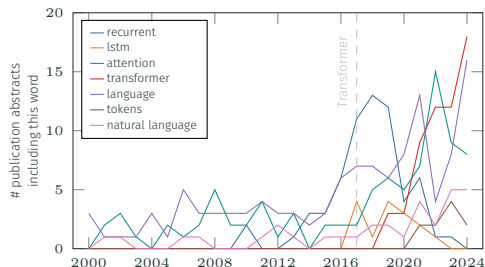
- Computer science applied to linguistics
- Interaction between computer and *written* human natural language

Music Information Retrieval (MIR)

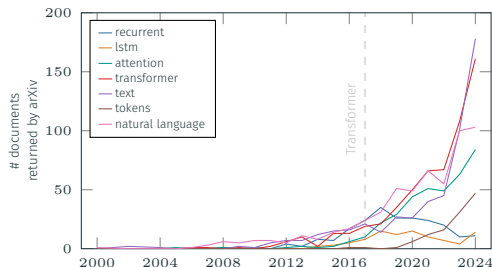
- Computer science applied to music
- Interaction between computer and (symbolic) music

NLP tools in symbolic MIR research

Number of ISMIR papers which include NLP-related terms in their abstract

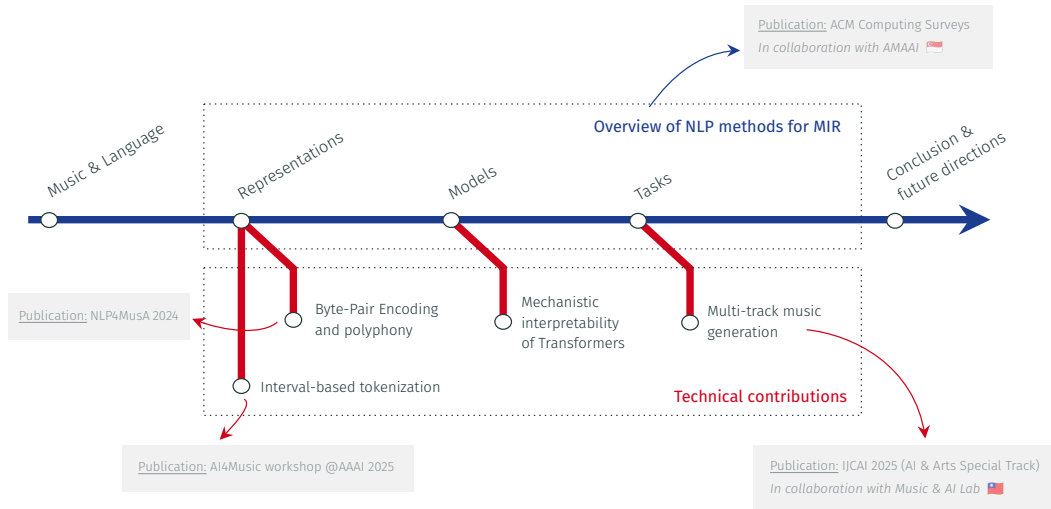


Number of MIR arXiv papers involving NLP-related terms

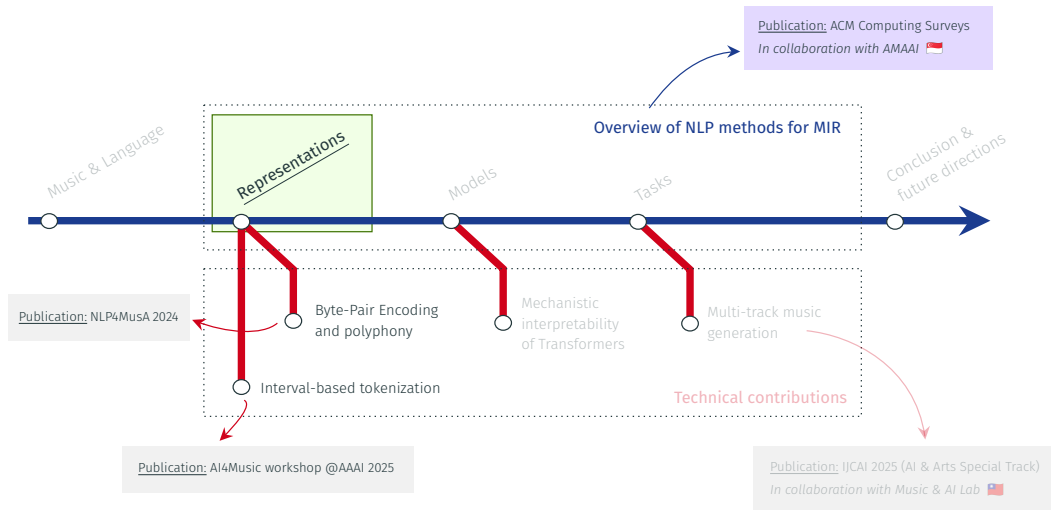


Le & al., *Natural Language Processing Methods for Symbolic Music Generation and Information Retrieval: A Survey*, ACM Computing Surveys 2025. In collaboration with AMAAI (Singapore University of Technology and Design).

Outline



Outline – Sequential representations



Tokenization

Representing a complex content (*e.g.* text, music) into a **sequence** of elements for computational processing.

Tokenization

Representing a complex content (*e.g.* text, music) into a **sequence** of elements for computational processing.

I love singing but it is exhausting

Word-level tokenization

→ <I>, <love>, <singing>, <but>, <it>, <is>, <exhausting>

Character-level tokenization

→ <I>, <_>, <l>, <o>, <v>, <e>, <_>, <s>, <i>, <n>, ...

Subword-level tokenization

→ <I>, <love>, <sing>, <_ing>, <but>, <it>, <is>, <exhaust>, <_ing>

What does a musical score encode?

Tempo Rhythm

101 ♩ = 80 a 2. *p*

Dynamics

Pitch

Articulations

Polyphony

Structure

The image displays a musical score for five instruments: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), and Violoncello/Contrabass (Vc. Cb.). The score is in common time (C) and begins at measure 101. A tempo marking of ♩ = 80 is present. The Oboe part features a melodic line starting in measure 103 with a dynamic of *p* and a second ending marked 'a 2.'. The Violin 1 and Violin 2 parts play a rhythmic pattern of eighth notes, with Violin 1 starting in measure 102 and Violin 2 in measure 103, both marked *p*. The Viola part plays a pizzicato (pizz.) pattern starting in measure 102, marked *p*, and then switches to arco (arco.) in measure 103. The Violoncello/Contrabass part plays a rhythmic pattern of eighth notes starting in measure 102, marked *p* and pizz. (pizz.). The score is annotated with various musical elements: Tempo (♩ = 80), Rhythm (indicated by note values), Dynamics (*p*), Pitch (indicated by clefs and note positions), Articulations (pizz., arco., a 2.), Polyphony (multiple parts playing simultaneously), and Structure (measure numbers 101, 102, 103).

What does a musical score encode?

Tempo

Rhythm

Dynamics

Pitch

Articulations

Polyphony

Structure

The image displays a musical score for five instruments: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), and Violoncello/Contrabass (Vc. Cb.). The score is in common time (C) and begins at measure 101. A tempo marking of $\text{♩} = 80$ is present. The Oboe part features a half note with a first ending bracket labeled 'a 2.' and a dynamic marking of *p*. The Violin 1 part has a blue square highlighting a quarter note with a dynamic marking of *p*. The Violin 2 part has a blue square highlighting a quarter note with a dynamic marking of *p*. The Viola part has a blue square highlighting a quarter note with a dynamic marking of *p* and an articulation marking of 'pizz.'. The Violoncello/Contrabass part has a blue square highlighting a quarter note with a dynamic marking of *p* and an articulation marking of 'pizz.'. The score is annotated with various musical elements: Tempo ($\text{♩} = 80$), Rhythm (note values), Dynamics (*p*), Pitch (note heads), Articulations ('pizz.', 'arco.', 'a 2.'), Polyphony (multiple staves), and Structure (measure numbers).

What does a musical score encode?

Tempo

Rhythm

Dynamics

Pitch

Articulations

Polyphony

Structure

The image displays a musical score for a string quartet and oboe, starting at measure 101. The tempo is marked as $\text{♩} = 80$. The score includes five staves: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), and Violoncello/Contrabass (Vc. Cb.).

Annotations and features include:

- Tempo:** $\text{♩} = 80$ is indicated at the top.
- Rhythm:** A red box highlights a half note in the Oboe staff, labeled "a 2.".
- Dynamics:** The word *p* (piano) is written below the first measures of the Oboe, Violin 1, Violin 2, and Viola staves.
- Pitch:** A blue box highlights a quarter note in the Viola staff, labeled "pizz." (pizzicato).
- Articulations:** A red box highlights a quarter note in the Viola staff, labeled "arco." (arco).
- Polyphony:** The word *pizz.* is written below the first measure of the Violoncello/Contrabass staff.
- Structure:** The word *pizz.* is written below the first measure of the Violoncello/Contrabass staff.

What does a musical score encode?

Tempo

Rhythm

Dynamics

Pitch

Articulations

Polyphony

Structure

The image displays a musical score for a string quartet and oboe, starting at measure 101. The score is annotated with various musical concepts:

- Tempo:** Indicated by a green box with a quarter note and the text "♩ = 80".
- Rhythm:** Indicated by a red box with a half note and the text "a 2." above it.
- Dynamics:** Indicated by a green box with the text "p" (piano) below the first violin part.
- Pitch:** Indicated by a blue box with the text "pizz." (pizzicato) below the viola part.
- Articulations:** Indicated by a red box with the text "arco." (arco) below the viola part.
- Polyphony:** Indicated by a blue box with the text "pizz." (pizzicato) below the cello part.
- Structure:** Indicated by a red box with the text "a 2." above the oboe part.

What does a musical score encode?

The image displays a musical score for 'The Swan' from Tchaikovsky's Swan Lake, featuring five staves: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), and Violoncello/Double Bass (Vc. Cb.). The score is annotated with various musical concepts:

- Tempo:** Indicated by a green box with the notation $\text{♩} = 80$.
- Rhythm:** Indicated by a red box with the notation *a 2.* (alla breve).
- Dynamics:** Indicated by a green box with the notation *p* (piano).
- Pitch:** Indicated by a blue box with the notation *pizz.* (pizzicato) and *arco.* (arco).
- Polyphony:** Indicated by a green box with the notation *pizz.* (pizzicato).
- Structure:** Indicated by a red box with the notation *a 2.* (alla breve).

The score is written in 2/4 time and features a variety of musical notations, including notes, rests, and articulations. The Oboe part begins with a whole note, followed by a half note. The Violin 1 and Violin 2 parts feature a series of eighth notes. The Viola part begins with a half note, followed by a half note. The Violoncello/Double Bass part features a series of eighth notes. The score is annotated with various musical concepts, including Tempo, Rhythm, Dynamics, Pitch, Polyphony, and Structure.

What does a musical score encode?

Tempo

Rhythm

Dynamics

Pitch

Articulations

Polyphony

Structure

The image displays a musical score for a string quartet, featuring five staves: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), and Violoncello/Contrabass (Vc. Cb.). The score is annotated with various musical elements:

- Tempo:** A green box indicates a tempo of $\text{♩} = 80$.
- Rhythm:** A pink box highlights a measure in the Oboe staff with a second ending bracket labeled "a 2.".
- Dynamics:** A grey box highlights the dynamic marking *p* (piano) in the Violin 1 staff.
- Pitch:** A blue box highlights a note in the Viola staff.
- Articulations:** A grey box highlights the articulation marking *pizz.* (pizzicato) in the Viola staff.
- Polyphony:** An orange box highlights the Violoncello/Contrabass staff.
- Structure:** A pink box highlights a measure in the Violoncello/Contrabass staff.

What does a musical score encode?

Tempo

Rhythm

Dynamics

Pitch

Articulations

Polyphony

Structure

The image displays a musical score for a string quartet, featuring five staves: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), and Violoncello/Double Bass (Vc. Cb.). The score is annotated with various musical elements:

- Tempo:** A green box indicates a tempo of $\text{♩} = 80$.
- Rhythm:** A pink box highlights a measure in the Oboe staff with a half note and a second ending bracket labeled "a 2.".
- Dynamics:** A grey box highlights the dynamic marking *p* (piano) in the Violin 1 staff.
- Pitch:** A blue box highlights a note in the Viola staff.
- Articulations:** A light blue box highlights a note in the Violin 2 staff.
- Polyphony:** An orange box highlights a measure in the Violoncello/Double Bass staff.
- Structure:** A pink box highlights a measure in the Violoncello/Double Bass staff.

Other annotations include "101" in a pink box, "pizz." (pizzicato) in a blue box, and "arco." (arco) in a blue box.

What does a musical score encode?

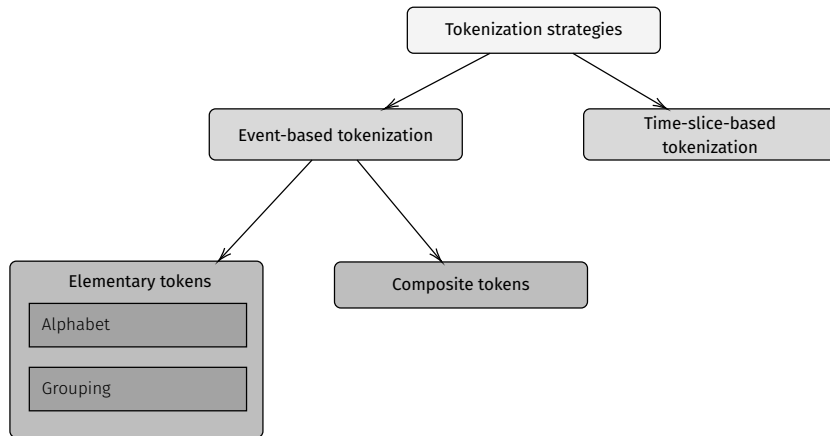
The image displays a musical score for five instruments: Oboe (Ob.), Violin 1 (Vln.1), Violin 2 (Vln.2), Viola (Vla.), and Violoncello/Contrabasso (Vc. Cb.). The score is annotated with various musical elements:

- Tempo:** Indicated by a green box with a quarter note and the text "♩ = 80".
- Rhythm:** Indicated by a pink box.
- Dynamics:** Indicated by a yellow box, with annotations like *p* (piano) and *a 2.* (second ending).
- Pitch:** Indicated by a blue box.
- Articulations:** Indicated by a light blue box.
- Polyphony:** Indicated by an orange box.
- Structure:** Indicated by a pink box.

Specific annotations in the score include:

- Ob.: *a 2.* (second ending), *p* (piano).
- Vln.1: *p* (piano).
- Vln.2: *p* (piano).
- Vla.: *pizz.* (pizzicato), *p* (piano), *arco.* (arco).
- Vc. Cb.: *pizz.* (pizzicato).

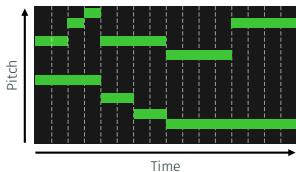
A taxonomy of tokenization strategies for symbolic music



How to tokenize music?



Pianoroll slices



MIDI-VAE [Brunner & al. 2018]

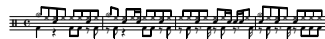
Harmony Transformer [Chen & al. 2019]

RNBert [Sailor 2024]



D5, __, E5, F5, D5, __, __, __, C5, __, __, __, E5
A4, __, __, G4, __, F4, __, E4, __, __, __, E4
C4, __, __, B3, __, __, __, G3, __, __, __, A3
F3, __, D3, __, G3, __, __, C2, __, __, C#2
1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4, 1
0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 0

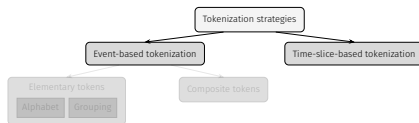
DeepBach [Hadjeres & al. 2017]



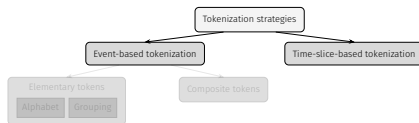
1	0--0--	17	---0--	33	---0--	49	0-0--
2	-----	18	-----	34	-----	50	-----
3	---0--	19	---00--	35	---00--	51	---00--
4	-----	20	-----	36	-----	52	-----
5	---0-0--	21	---0-0--	37	---0-0--	53	---0-0--
6	-----	22	-----	38	-----	54	-----
7	---0-0--	23	---0-0--	39	---0-0--	55	---0-0--
8	-----	24	---0-0--	40	---0-0--	56	---0-0--
9	---00--	25	---00--	41	---0-0--	57	---0-0--
10	-----	26	-----	42	-----	58	-----
11	---00--	27	---00--	43	---00--	59	---00--
12	-----	28	-----	44	---0--	60	-----
13	---0-0--	29	---0-0--	45	---0--	61	---00--
14	-----	30	-----	46	---0--	62	-----
15	---00--	31	---00--	47	---00--	63	---00--
16	-----	32	-----	48	---0--	64	-----

Drums tokenization [Zhang & al. 2023]

How to tokenize music?

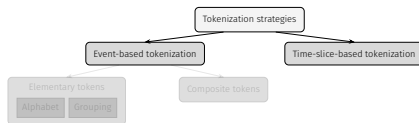


How to tokenize music?

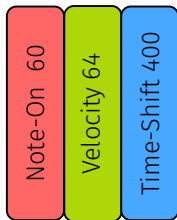
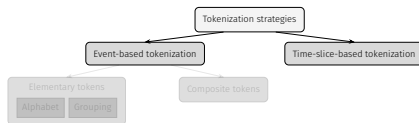


Note-On 60

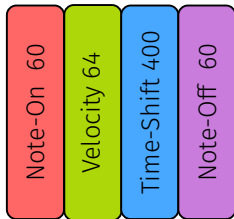
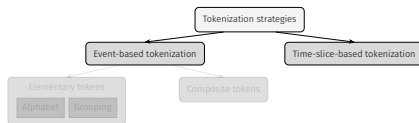
How to tokenize music?



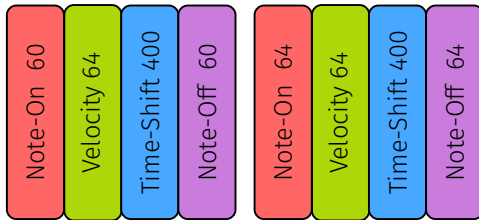
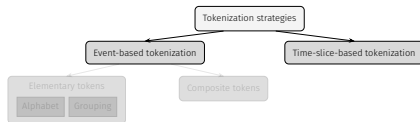
How to tokenize music?



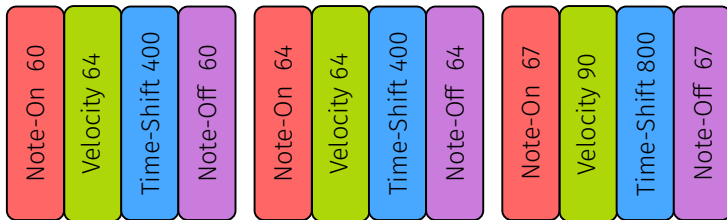
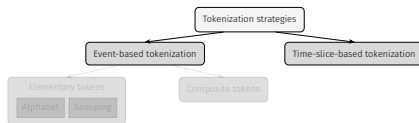
How to tokenize music?



How to tokenize music?

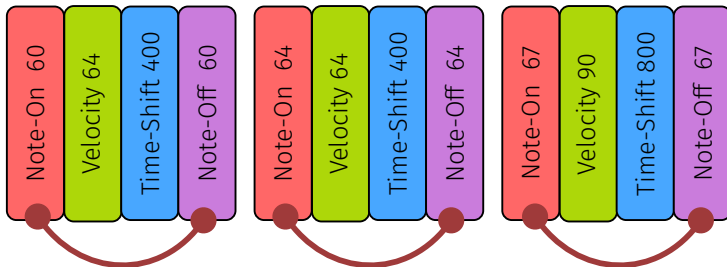
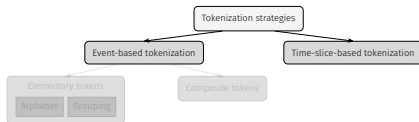


How to tokenize music?

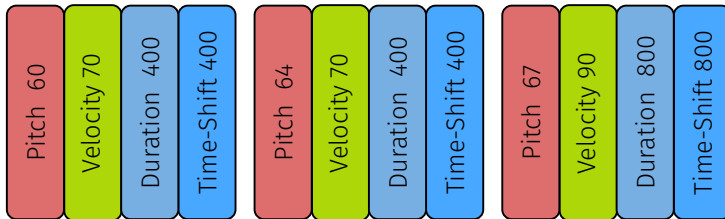
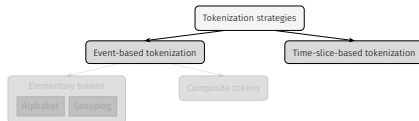


MIDI-like: Huang & al., *Music Transformer: Generating Music with Long-term Structure*, ICLR 2019

How to tokenize music?

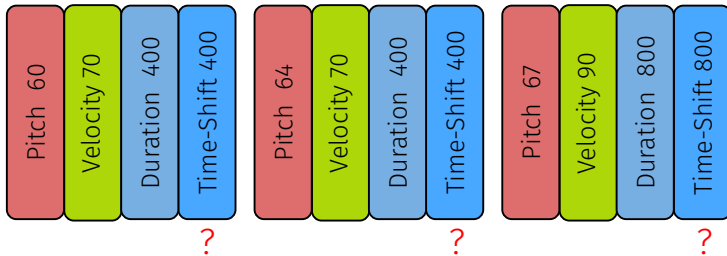
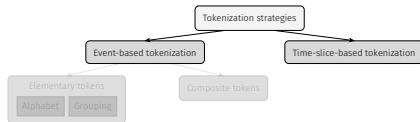


How to tokenize music?

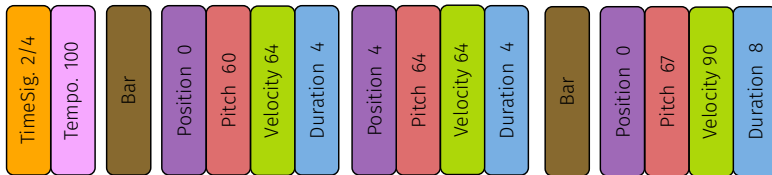
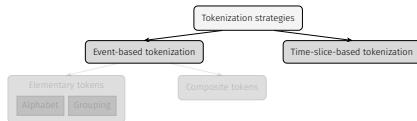


Structured: Hadjares & al., *The Piano Inpainting Application*, 2021

How to tokenize music?

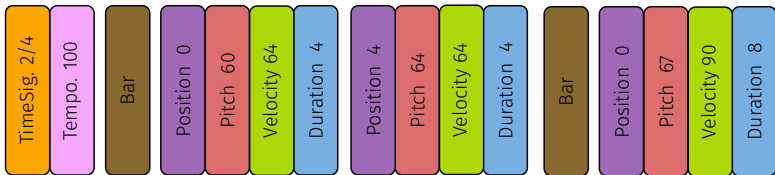
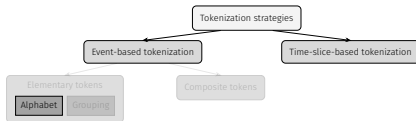


How to tokenize music?

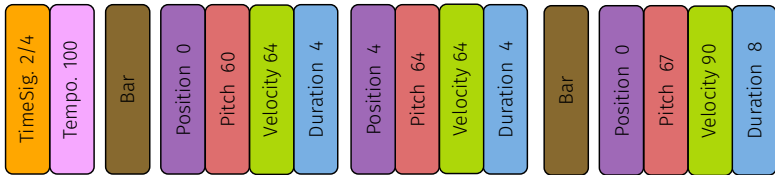
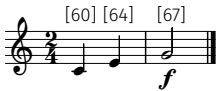
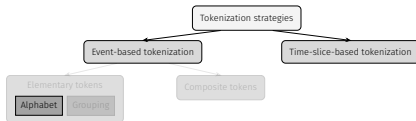


REMI: Huang & al., *Pop Music Transformer*, MM 2020

How to tokenize music?

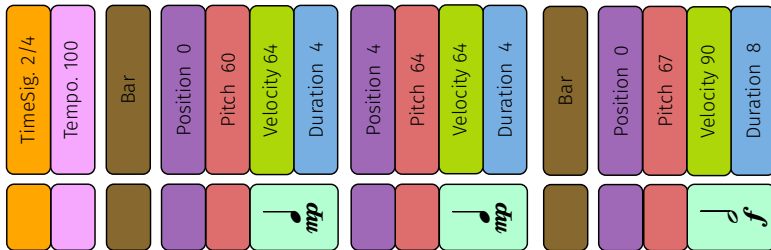
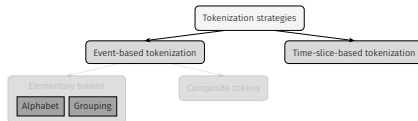


How to tokenize music?



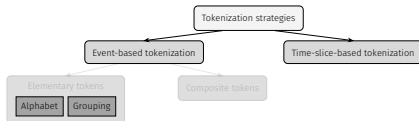
Grouping tokens together?

How to tokenize music?

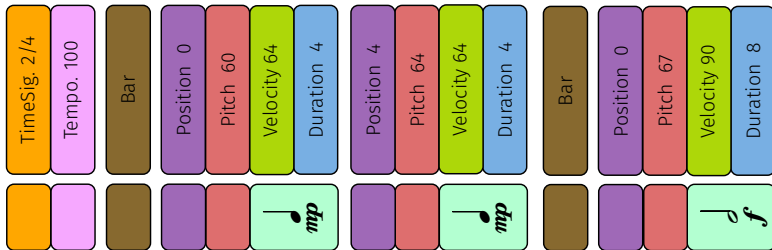


Rule-based ; n-gram: Conklin & al., 1995; Byte-Pair Encoding: Fradet & al., 2023

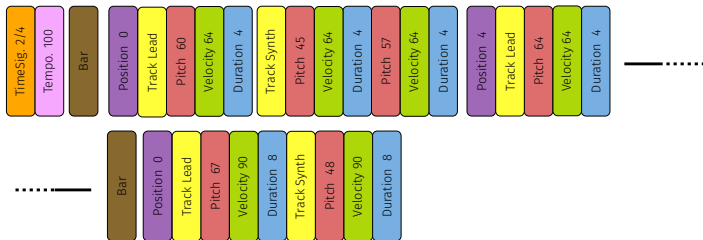
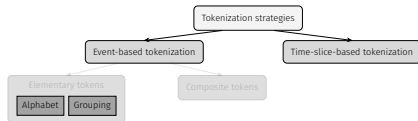
How to tokenize music?



Multi-track?

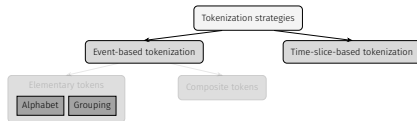


How to tokenize music?



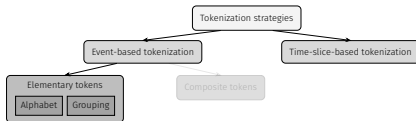
REMI+: von Rutte & al., FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control, ICLR 2023

How to tokenize music?

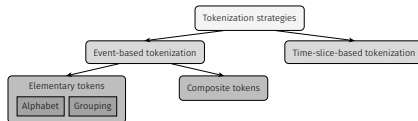


Sequence length?

How to tokenize music?



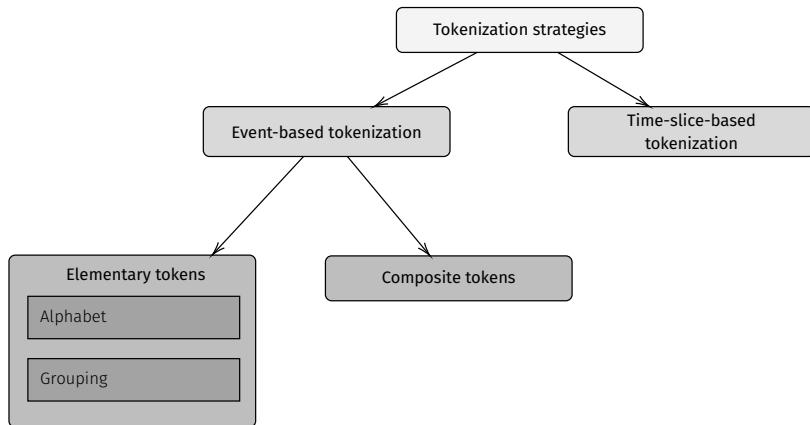
How to tokenize music?



TimeSig. 2/4	TimeSig. 2/4	TimeSig. 2/4	TimeSig. 2/4	TimeSig. 2/4	TimeSig. 2/4
Tempo. 100	Tempo. 100	Tempo. 100	Tempo. 100	Tempo. 100	Tempo. 100
Bar 1	Bar 1	Bar 1	Bar 1	Bar 2	Bar 2
Position 0	Position 0	Position 0	Position 4	Position 0	Position 0
Track Lead	Track Synth	Track Synth	Track Lead	Track Lead	Track Synth
Pitch 60	Pitch 45	Pitch 57	Pitch 64	Pitch 67	Pitch 48
Velocity 64	Velocity 64	Velocity 64	Velocity 64	Velocity 90	Velocity 90
Duration 4	Duration 4	Duration 4	Duration 4	Duration 8	Duration 8

Octuple: Zeng & al., *MusicBERT: Symbolic Music Understanding with Large-Scale Pre-Training*, ACL 2021

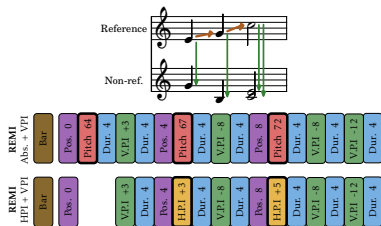
A taxonomy of tokenization strategies for symbolic music



Le & al., *Natural Language Processing Methods for Symbolic Music Generation and Information Retrieval: A Survey*, ACM Computing Surveys 2025. In collaboration with AMAAI (Singapore University of Technology and Design).

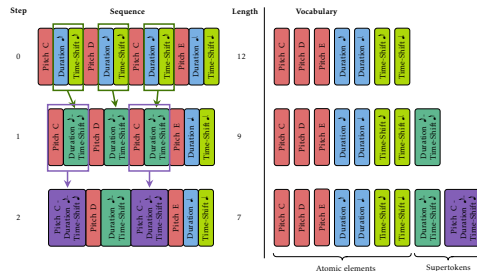
Technical contributions – Tokenization expressiveness

Alphabet: Interval-based tokenization



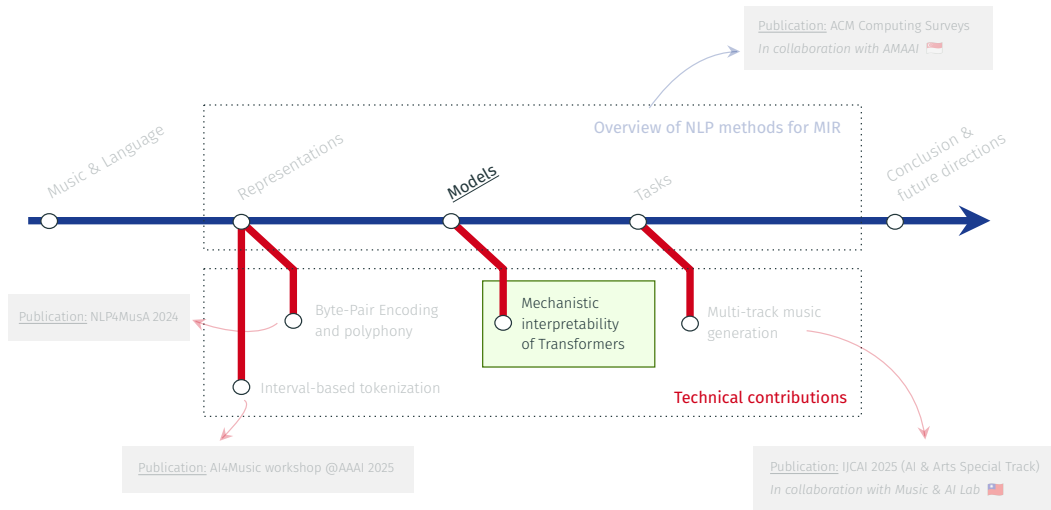
Le & al., *Evaluating Interval-based Tokenization for Pitch Representation in Symbolic Music Analysis*, AI4Music workshop at AAAI 2025

Grouping: Byte-pair encoding & polyphony

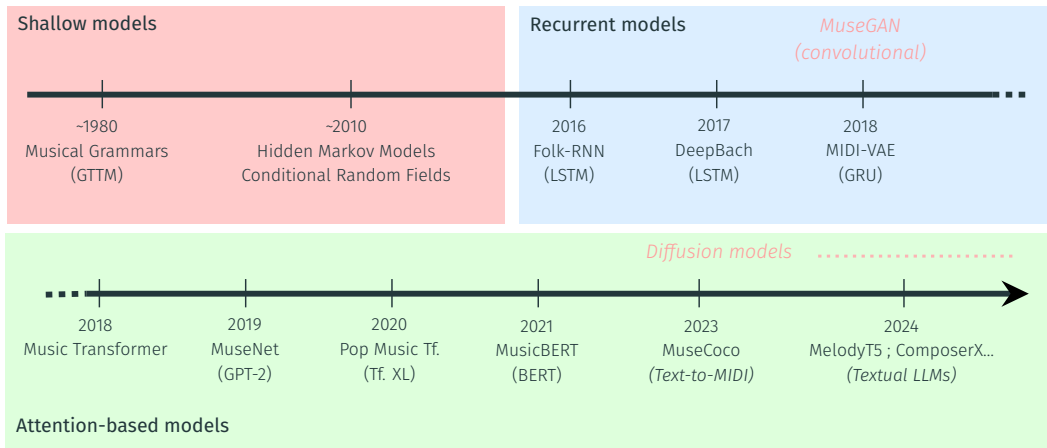


Le & al., *Analyzing Byte-Pair Encoding on Monophonic and Polyphonic Symbolic Music: A Focus on Musical Phrase Segmentation*, NLP4MusA 2024

Outline – Models



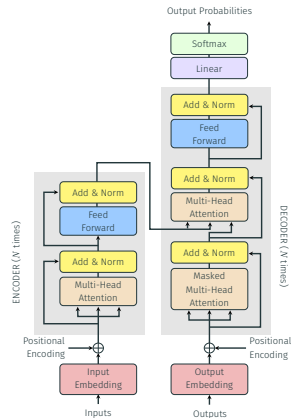
NLP models for symbolic music processing



Attention mechanism and Transformers

Attention mechanism

- Build new representations of an input token sequence
- Weight the importance of each token according to its relevance to the others



Vaswani & al., *Attention is All You Need*, 2017

Model interpretability (*or explainability*)

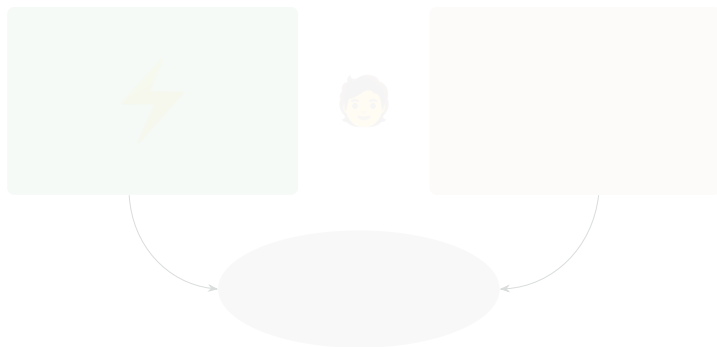
- Make a model's decision process transparent, faithful, and interpretable
- *Human-centered* explanations

In Music Information Retrieval:

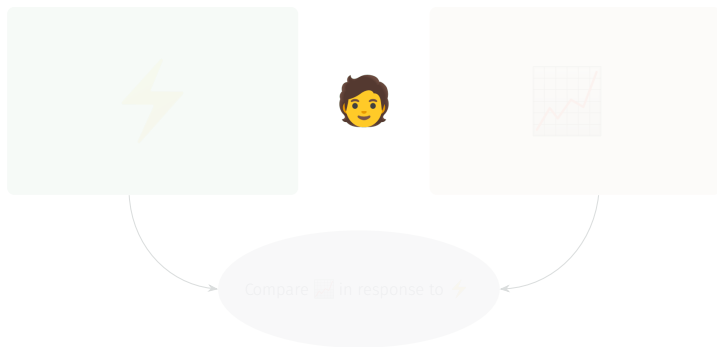
- **Chain-of-thoughts** for ABC Notation-based models [Zhou & al. 2024]
- Attention mechanism **visualization** in music generation [Huang & al. 2018]

⇒ Our approach: **Analyzing** the attention mechanism in the context of functional harmonic analysis

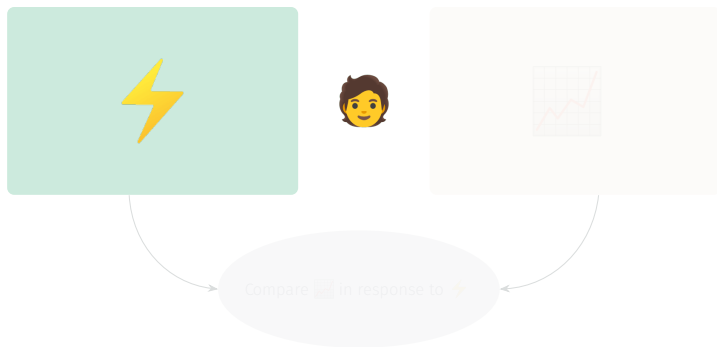
A methodology to interpret attention with regards to musical characteristics



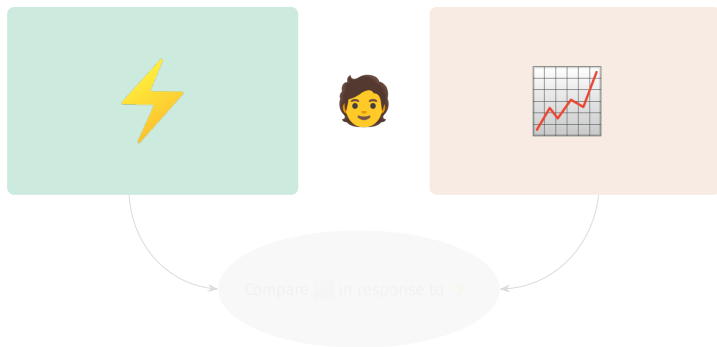
A methodology to interpret attention with regards to musical characteristics



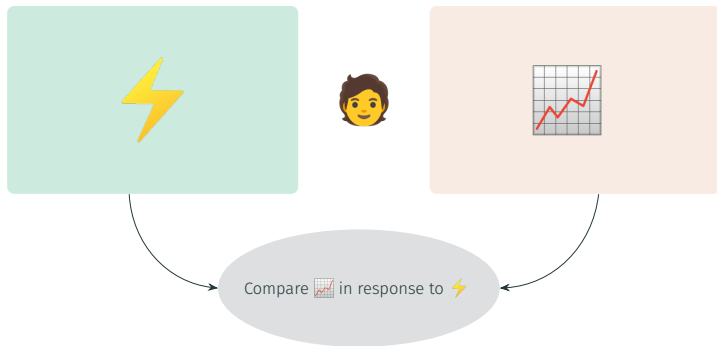
A methodology to interpret attention with regards to musical characteristics



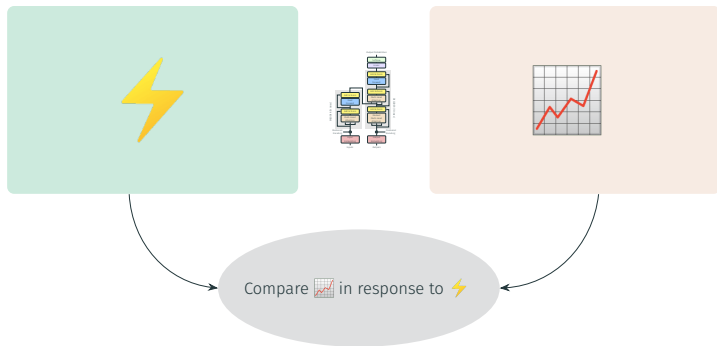
A methodology to interpret attention with regards to musical characteristics



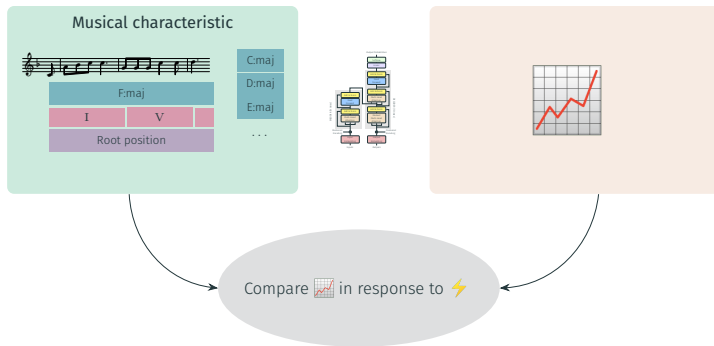
A methodology to interpret attention with regards to musical characteristics



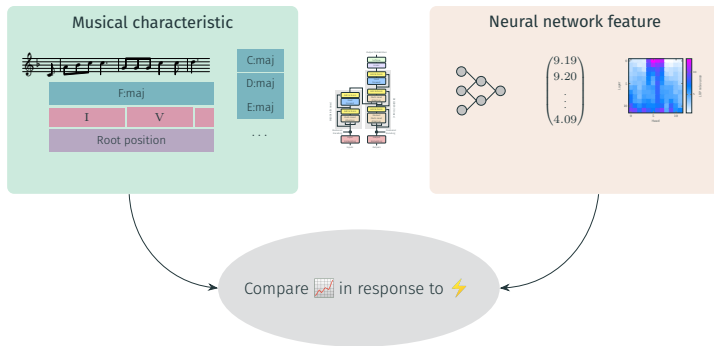
A methodology to interpret attention with regards to musical characteristics



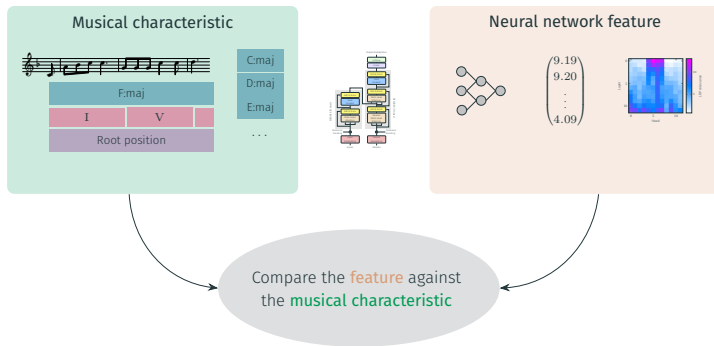
A methodology to interpret attention with regards to musical characteristics



A methodology to interpret attention with regards to musical characteristics



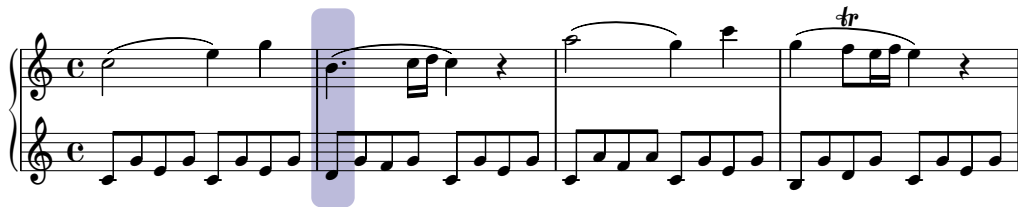
A methodology to interpret attention with regards to musical characteristics



Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang



Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

The image shows a musical score for Mozart's K. 545 (Lang-Lang) in C major, first movement. The score is in 3/4 time and consists of four measures. A vertical purple bar highlights the second measure. A green bracket under the first measure is labeled 'C maj' and 'Local key' with a dotted arrow. A green bracket under the last two measures is connected by a curved arrow.

Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

The image displays a musical score for Mozart's K. 545 (Lang-Lang) in C major, 2/4 time. The score is written for piano and features a treble and bass staff. A vertical purple bar highlights a specific measure. Below the score, several annotations are present:

- Local key:** A green dotted arrow points to the key signature (one sharp) and is labeled "C maj".
- Quality:** A blue dotted arrow points to the chord symbol G^7 and is labeled "Quality".
- Degree:** An orange dotted arrow points to the chord symbol $\leadsto V$ and is labeled "Degree".

Additional annotations include a green bracket under the first measure, a blue bracket under the second measure, and a green bracket under the fourth measure. A green curved arrow points from the first measure to the fourth measure. A blue curved arrow points from the first measure to the second measure. An orange curved arrow points from the second measure to the third measure. A green curved arrow points from the fourth measure to the fifth measure. A blue curved arrow points from the second measure to the third measure. An orange curved arrow points from the third measure to the fourth measure. A green curved arrow points from the fourth measure to the fifth measure.

Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

The image displays a musical score for Mozart's K. 545 (Lang-Lang) in C major, 2/4 time. The score is written for piano and features a treble and bass staff. A vertical purple bar highlights a specific measure. Below the score, several annotations are provided:

- Local key:** C maj (indicated by a green dotted arrow pointing to the key signature).
- Quality:** G⁷ (indicated by a blue dotted arrow pointing to the G7 chord).
- Degree:** V (indicated by an orange dotted arrow pointing to the V chord).
- Inversion:** 7 or 6/5 or 4/3 or 4/2 (indicated by a pink dotted arrow pointing to the inversion options).

Additional annotations include a green bracket under the first measure, a blue bracket under the second measure, and a green bracket under the third measure. A green curved arrow points from the first measure to the third measure.

Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

The image displays a musical score for Mozart's K. 545 (Lang-Lang) in C major, 2/4 time. The score is written for piano and features a treble and bass staff. A vertical purple bar highlights a specific measure. Below the score, several annotations are provided:

- Local key:** C maj (indicated by a green dotted arrow pointing to the key signature).
- Quality:** G⁷/D (indicated by a blue dotted arrow pointing to the highlighted measure).
- Degree:** V₃⁴ (indicated by an orange dotted arrow pointing to the highlighted measure).
- Inversion:** 7 or 6/5 or 4/3 or 4/2 (indicated by a pink dotted arrow pointing to the highlighted measure).

Additional annotations include a green bracket under the first measure, a green bracket under the last measure, and a green curved arrow pointing from the last measure back to the first measure.

Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

mozartk545-lang-lang

The image shows a musical score for the first movement of Mozart's Piano Sonata No. 11 (K. 545). The score is in C major, 2/4 time. The first two measures are highlighted with a green bracket and labeled "C maj". The third measure is highlighted with a purple vertical bar. Below the third measure, there is a pink arrow pointing from the G7/D chord to the V4/3 chord, with the text "G⁷/D ~ V⁴₃". To the right of this, there is a pink text "7 or 6/5 or 4/3 or 4/2". A green bracket under the last two measures is connected by a green curved arrow.

C maj

$G^7/D \rightsquigarrow V^4_3$

7 or $\frac{6}{5}$ or $\frac{4}{3}$ or $\frac{4}{2}$

Functional Harmonic Analysis (or Roman Numeral Analysis)

Is there a human attention mechanism somewhere?

$\approx$ “attention arrows” in text?

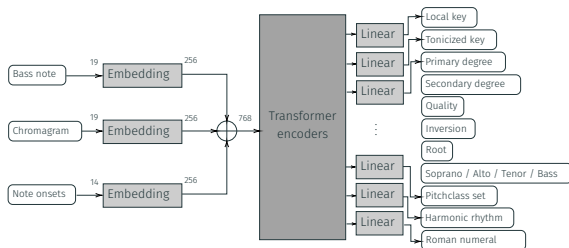


Automatic Functional Harmonic Analysis

Representation

- Time-slice-based tokenization: quantization at the 

Model architecture



Transformer encoders

- Stacked layers of multi-head self-attention¹
 - Bi-directional
 - 12 layers
 - 12 **attention heads** per layer

¹Vaswani & al., *Attention is All You Need*, 2017

Training

- End-to-end training
- 596 pieces, 28M tokens [Nápoles López & al., 2021]
- Era: baroque → early romantic
- Instruments: solo piano → string quartet

Performances (accuracy)

Model	Key	Degree	Quality	Inversion	Root	Global	# params
AugmentedNet v1.9.1	82.2	67.0	79.7	78.8	83.0	–	105k
RNBert	82.5	85.9	86.5	87.2	–	–	109M
Our model (mono-task)	79.5	81.7	90.0	89.7	92.0	87.5	–
Our model (multi-task)	81.7	84.3	90.9	90.3	92.8	88.6	94.9M

Understanding inner mechanisms in local key detection

modulation

John Rutter, *For the beauty of the earth*. (Performance: Taipei Male Choir)

The image displays a musical score for the piece 'For the beauty of the earth' by John Rutter, performed by the Taipei Male Choir. The score is presented in two systems, each with a vocal line and a piano accompaniment. The first system is highlighted in yellow. The second system is divided into two parts: the first part (measures 57-60) is highlighted in yellow, and the second part (measures 61-64) is highlighted in green. The key signature changes from B-flat major (two flats) to D major (two sharps) at measure 61, which is indicated by a double bar line and a key signature change symbol. The tempo is marked 'modulation'.

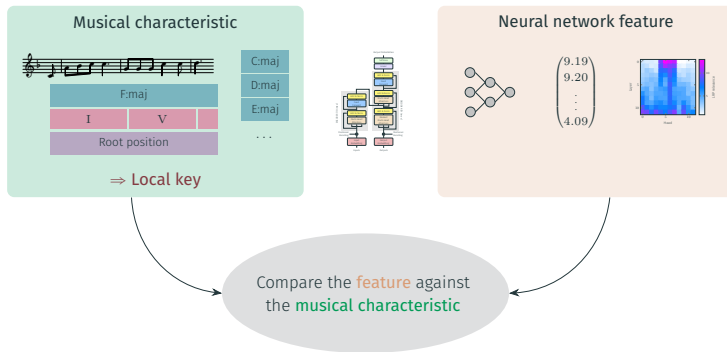
Understanding inner mechanisms in local key detection



Do the attention heads relevant for the annotation remain the same before and after the key change?

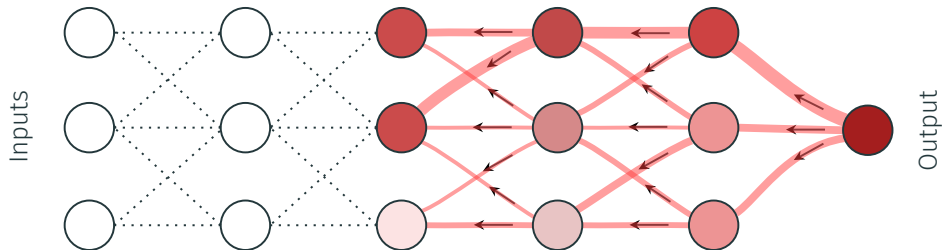
In the following, we will consider the model *only* trained on local key detection.

A methodology to interpret attention with regards to musical characteristics



Layer-wise relevance propagation (LRP)

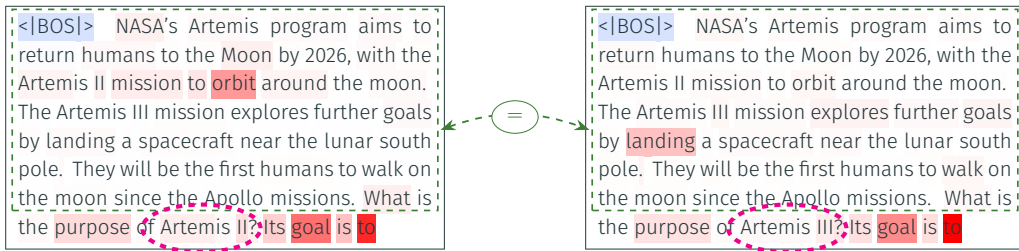
1. Consider a trained model on a 1-class classification task
2. Perform a forward pass and keep track of logits
3. LRP backward pass:



Montavon & al., *Layer-Wise Relevance Propagation: An Overview*, 2019

Layer-wise relevance propagation (LRP)

(In NLP) Model: Llama-3.2 (next word prediction)

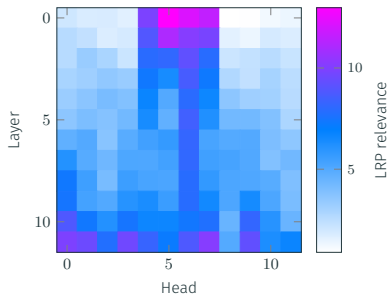


Achtibat & al., *AttnLRP: Attention-Aware Layer-Wise Relevance Propagation for Transformers*, 2024

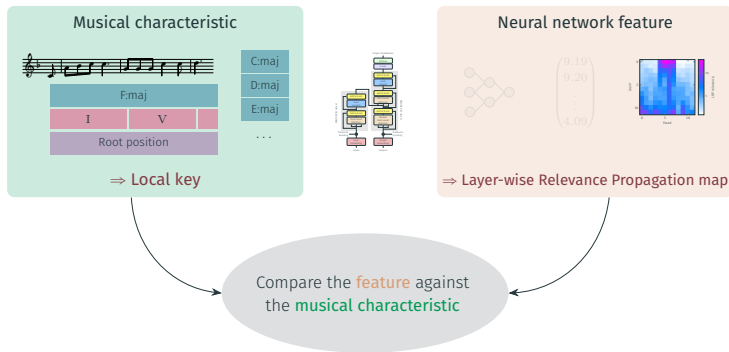
Layer-wise relevance propagation (LRP)

What if back-propagation is not performed up to the initial tokens?

LRP map: contribution of each attention head to a prediction



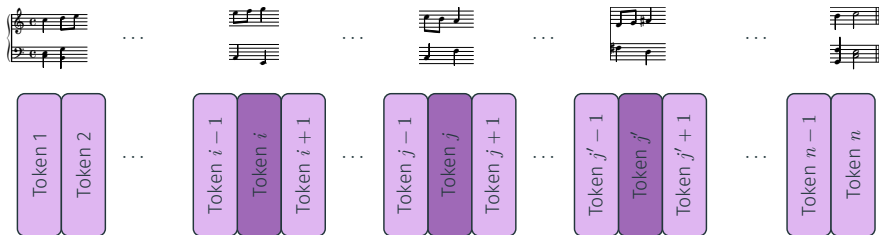
A methodology to interpret attention with regards to musical characteristics



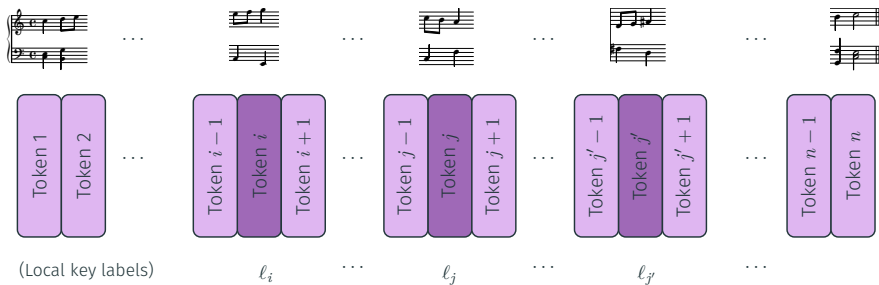
Attention head relevance for local key detection



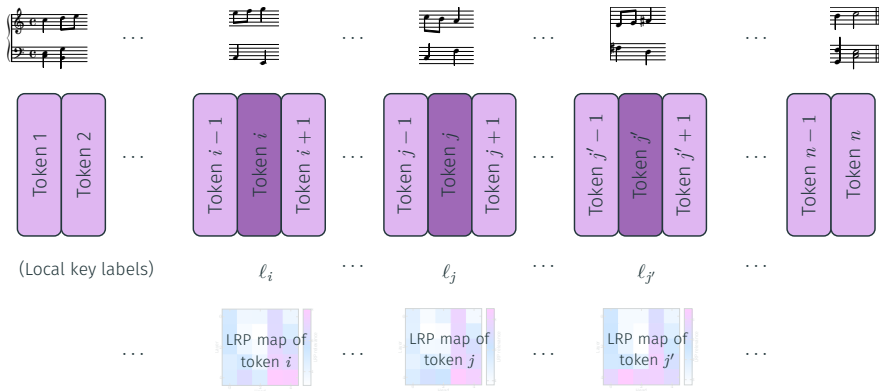
Attention head relevance for local key detection



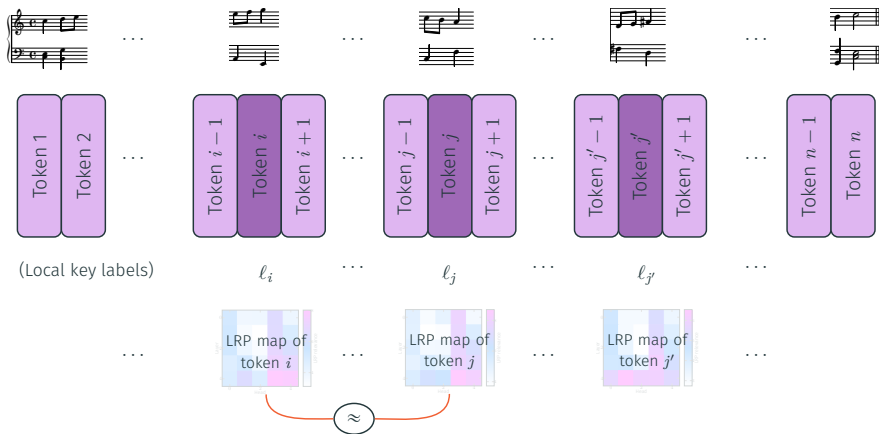
Attention head relevance for local key detection



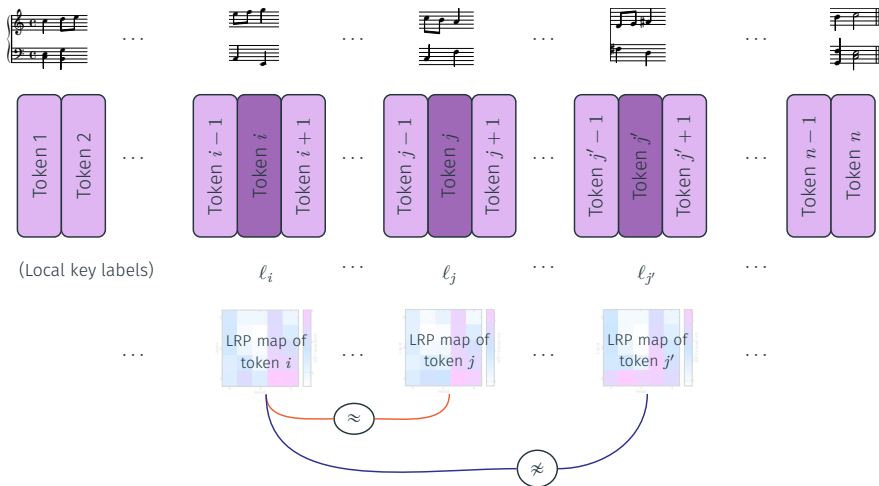
Attention head relevance for local key detection



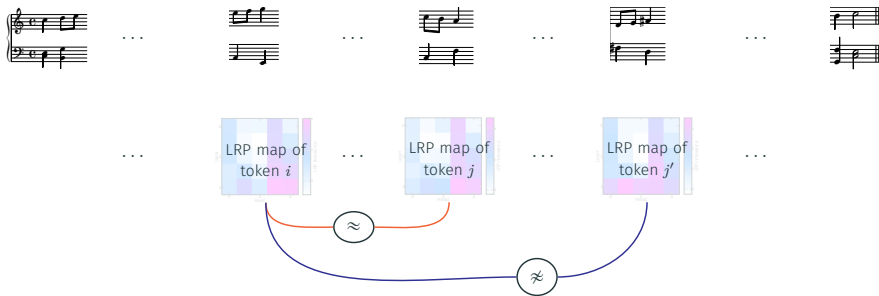
Attention head relevance for local key detection



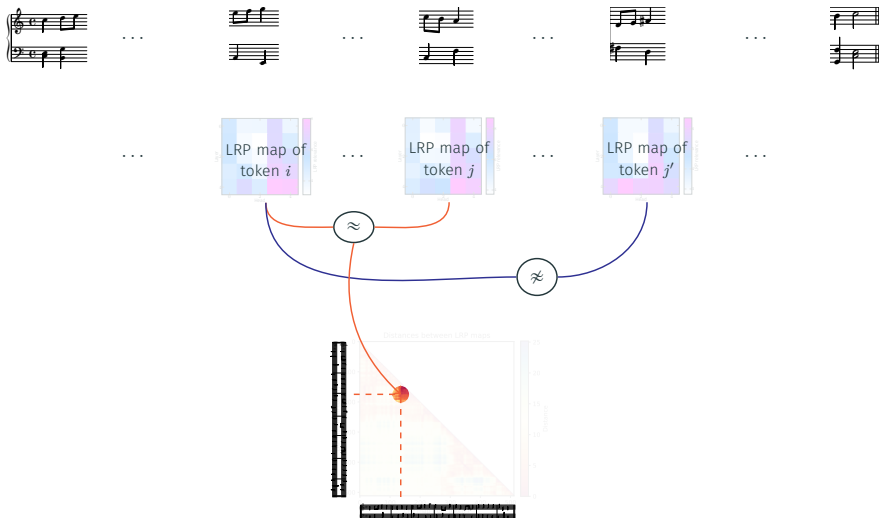
Attention head relevance for local key detection



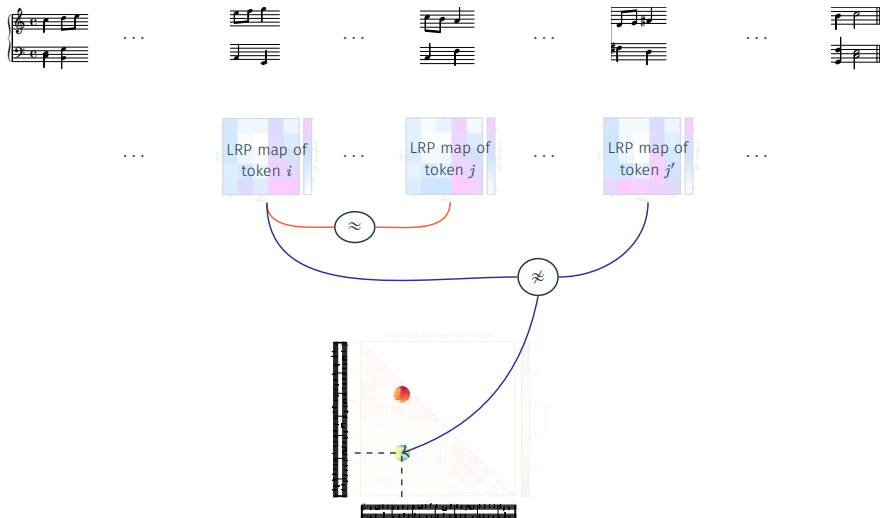
Attention head relevance for local key detection



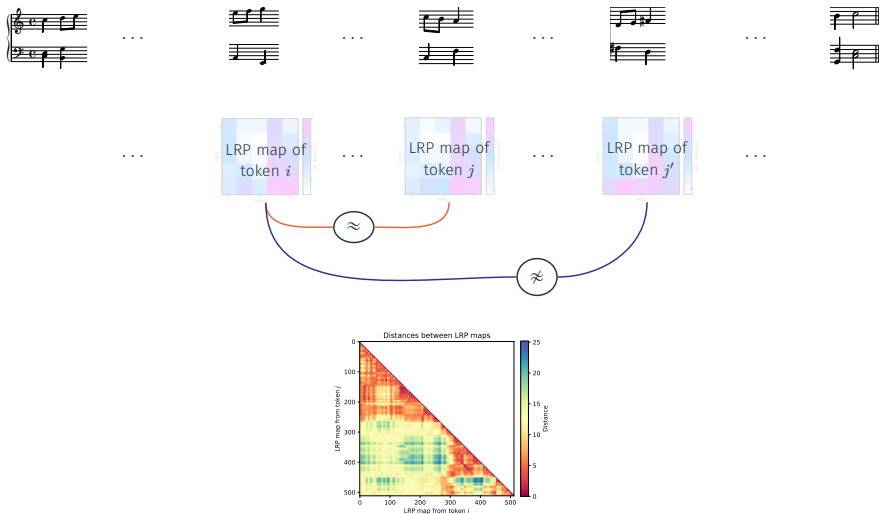
Attention head relevance for local key detection



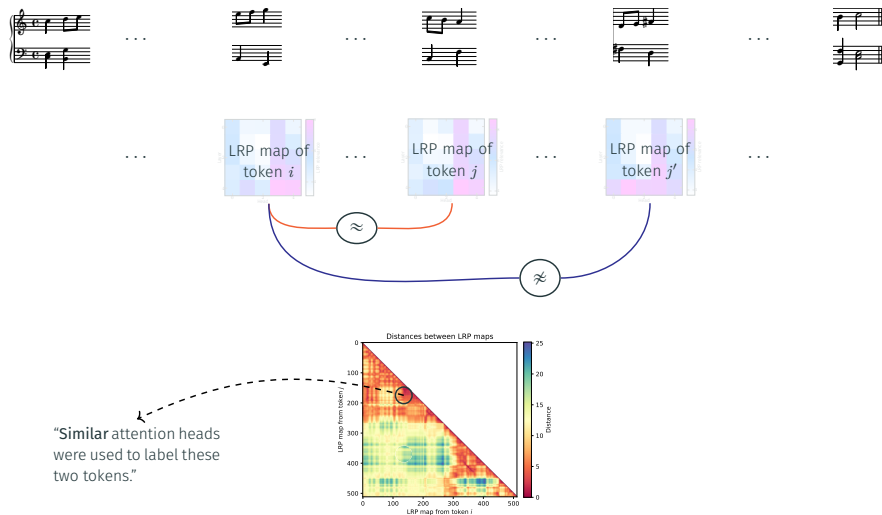
Attention head relevance for local key detection



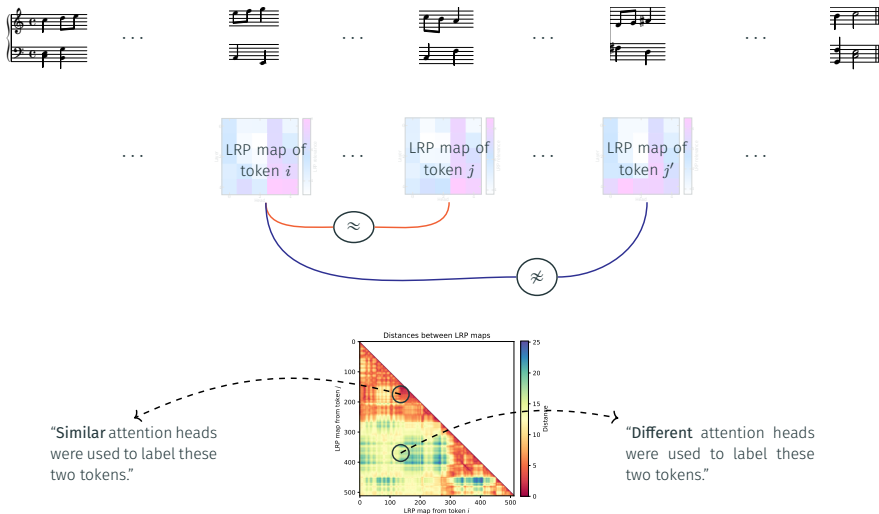
Attention head relevance for local key detection



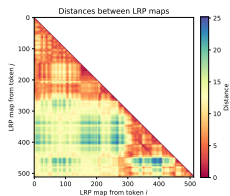
Attention head relevance for local key detection



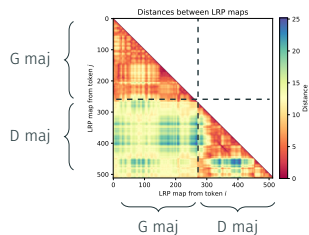
Attention head relevance for local key detection



Relating distances to labels

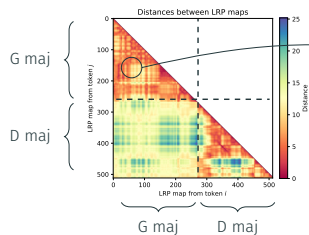


Relating distances to labels



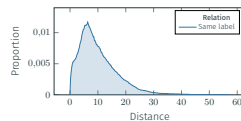
Attention head relevance for local key detection

Relating distances to labels



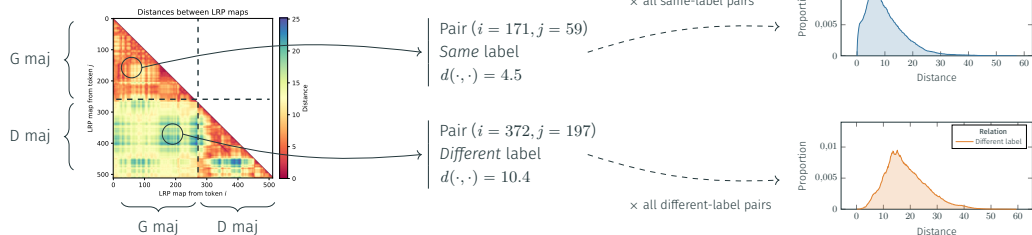
Pair ($i = 171, j = 59$)
Same label
 $d(\cdot, \cdot) = 4.5$

× all same-label pairs



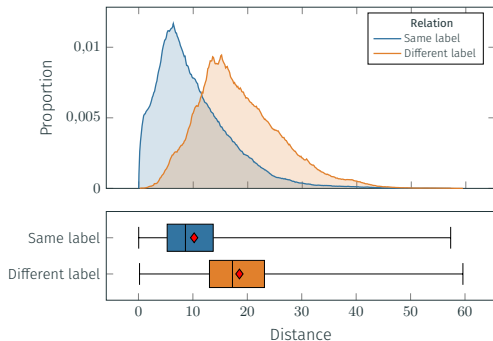
Attention head relevance for local key detection

Relating distances to labels



⇒ Comparing two distributions

Attention head relevance for local key detection



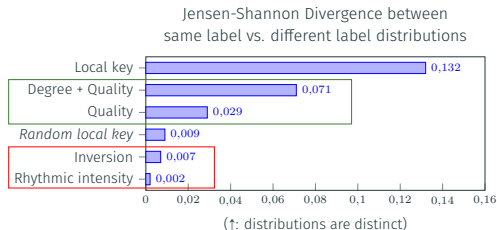
- The distributions between same vs. different labels are distinct (▲ | ▲)
- ⇒ The relevant attention heads are different when changing keys.
- Distances for same label pairs are lower than different label pairs (◆)

The attention heads relevant for the annotation of the *same local key* are *more similar* than for the annotation of different local keys.

Attention head relevance for local key detection

Comparing with other features

What if we consider other labels beyond local key?



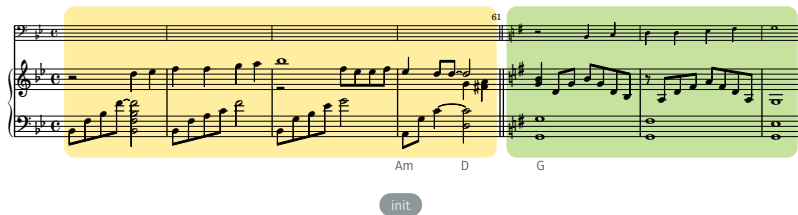
- Other harmonic features are also present in the model.
- Less relevant features for local key detection show less signal.

A model *only* trained on local key detection can capture *human-relevant* harmonic features it has not been trained on.

Attention head relevance for local key detection

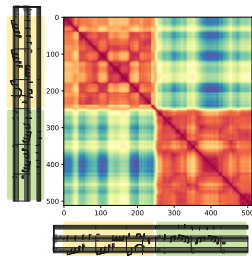
What leads the model to “think” that a modulation has occurred?

Original modulation (B♭ maj → G maj)



The image displays a musical score for John Rutter's "For the beauty of the earth". The score is divided into two sections: a yellow-shaded section on the left and a green-shaded section on the right. The yellow section is labeled with "Am" and "D" below it, and the green section is labeled with "G" below it. A small "init" button is located below the yellow section. The score is written for a piano, with a treble and bass staff. The modulation from B-flat major to G major is indicated by the change in the key signature and the labels below the score.

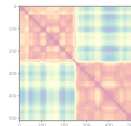
Distance between LRP maps



John Rutter, *For the beauty of the earth*.

Attention head relevance for local key detection

Initial piece

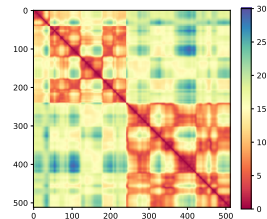


Keep the key signature

($\Rightarrow$ Modulation
Bb maj $\rightarrow$ G min)

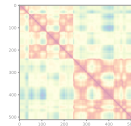


keep-key-signature



Attention head relevance for local key detection

Last step:
Key signature kept

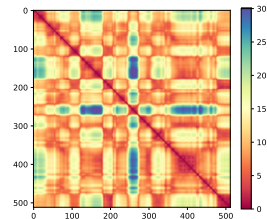


Transpose the modulated part to the initial key

(B $\flat$ maj)

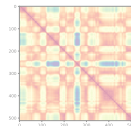


transpose



Attention head relevance for local key detection

Last step:
Key signature kept
+ Transposition

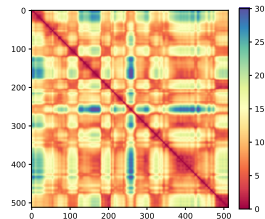


Remove the leading tone of the initial modulation

(F# in G maj)



no-fis



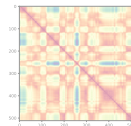
Attention head relevance for local key detection

Last step:

Key signature kept

+ Transposition

+ No leading tone



Re-write the initial modulation preparation

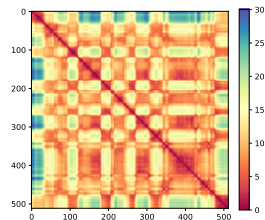
ii $\rightarrow$ V $\rightarrow$ I

G major (Am $\rightarrow$ D $\rightarrow$ G)

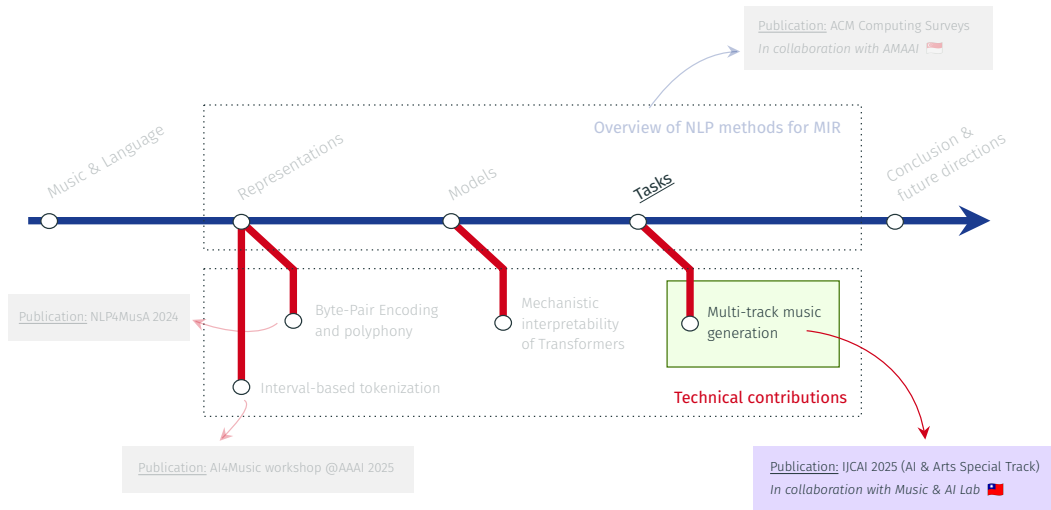
to B $\flat$ major (Cm $\rightarrow$ F $\rightarrow$ B $\flat$)



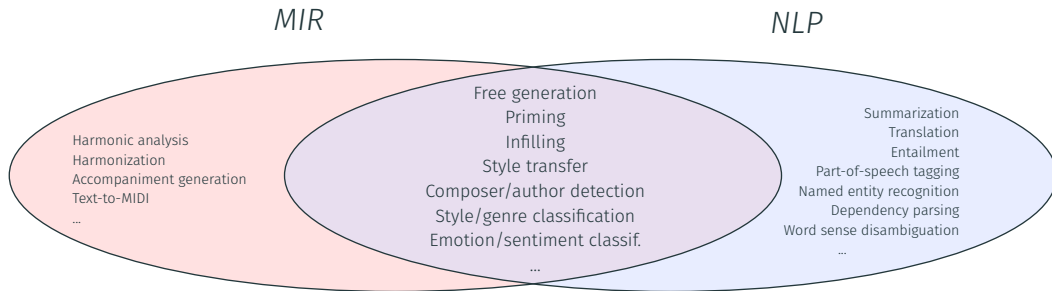
no-preparation



Outline – Tasks



Tasks in NLP and MIR



Re-orchestration: instrumentation, texture, melody

Instrumentation

Symphonic orchestra

Flauti.

Oboi.

Clarineti in B.

Fagotti.

Corni in Es.

Trombe in C.

Timpani in C.G.

Violino I.

Violino II.

Viola.

Violoncello.

Basso.

String orchestra

Piano - violin duet

Violine.

Allegro con brio. (♩ = 108.)

Pianoforte.

Piano solo

A musical score for the song 'The Rose Tree'. It features a piano accompaniment and a vocal line. The piano part is in 3/4 time, with a key signature of one flat (B-flat). The vocal line is in 3/4 time, with a key signature of one flat. The piano part includes a section marked 'ff' (fortissimo) and 'Pizzicato' (pizzicato). The vocal line includes a section marked 'ff' (fortissimo) and 'Pizzicato' (pizzicato). The score is written for a piano and voice.

Organ solo

Allegro con brio

Manuals

ff

Pedal

L.v. Beethoven, *Symphony No. 5*, Mvnt. 1, Bars 1–5.

Orchestral texture

Melodic fidelity

Re-orchestration: instrumentation, texture, melody

Instrumentation

Orchestral texture



Re-orchestration: instrumentation, texture, melody

Instrumentation

Orchestral texture

Melodic fidelity

朋友 (Wakin Chau)



いつも何度でも (from Spirited Away - Joe Hisaishi)



Ritournelle



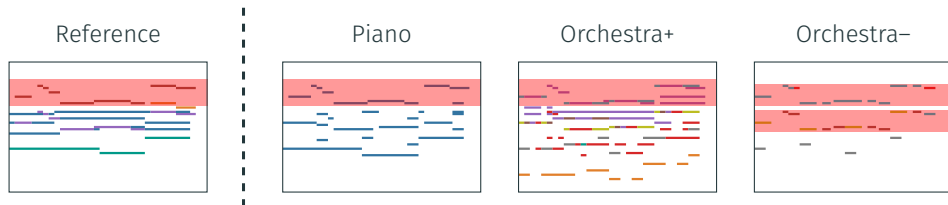
你被寫在我的歌裡 (sodagreen)



("Pachelbel canon"-like chord progression: I – V – vi – iii – IV – I – (ii) – V)

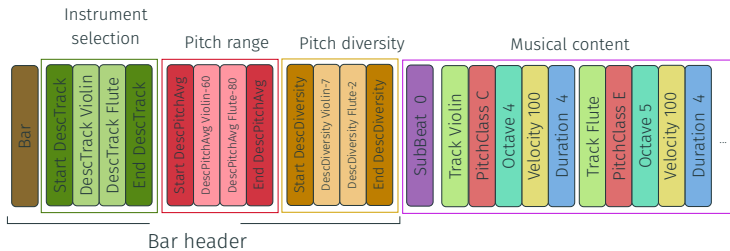
METEOR:

Melody-aware Texture-controllable Symbolic Music Re-orchestration via Transformer VAE



Le & Yang, *METEOR: Melody-aware Texture-controllable Symbolic Orchestral Music Generation via Transformer VAE*, IJCAI 2025 (AI & Arts Special Track). **In collaboration with Music & AI Lab (National Taiwan University).**

Tokenization & Textural controls



Bar-level

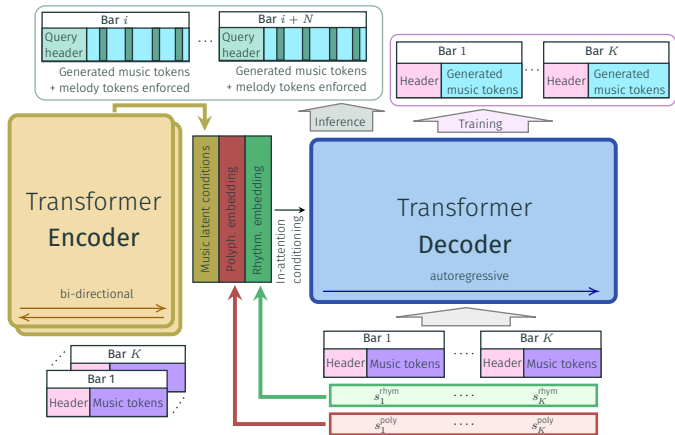
- Rhythmic intensity
- Polyphonicity

Bar+track-level

- Average pitch
- Pitch diversity

⇒ Derived statistically from the dataset

Model architecture



Wu & Yang, *MuseMorphose: Full-Song and Fine-Grained Piano Music Style Transfer with One Transformer VAE*, 2022

Melodic fidelity: inference guidance

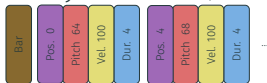
Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)
2. Choice of melodic instruments (e.g. `<Track-trumpet>`)
3. During inference

Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)

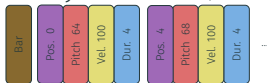


2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference

Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



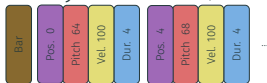
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



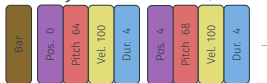
2. Choice of melodic instruments (e.g. `<Track-trumpet>`)
3. During inference



Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



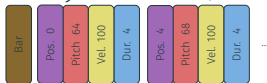
2. Choice of melodic instruments (e.g. `<Track-trumpet>`)
3. During inference



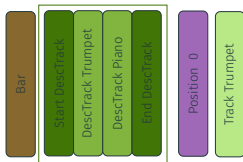
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



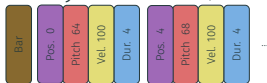
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



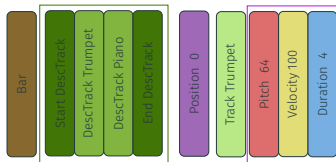
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



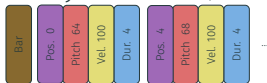
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



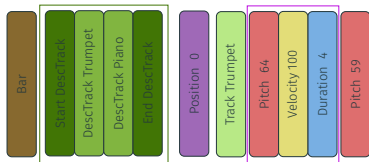
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



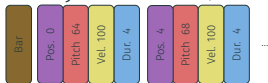
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



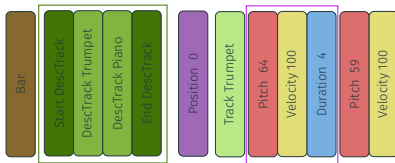
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



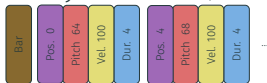
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



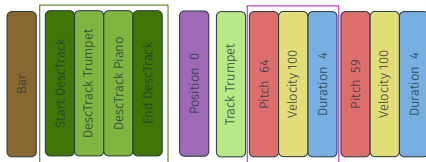
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



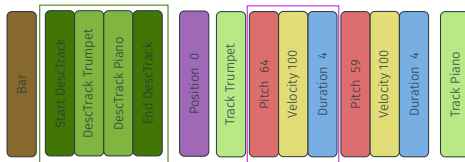
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



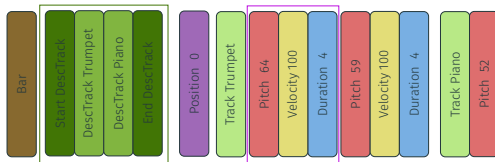
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



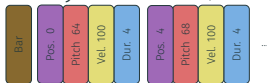
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



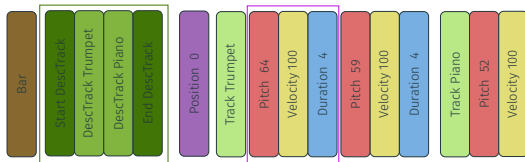
Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



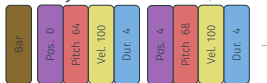
2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



Melodic fidelity: inference guidance

Melody enforcing **during inference**.

1. Melody detection (bar-wise / instrument-wise skyline)



2. Choice of melodic instruments (e.g. <Track-trumpet>)
3. During inference



Lead sheet orchestration... for free?

Lead sheet: melody + chords

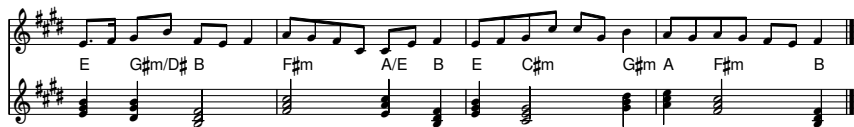


Lead sheet orchestration... for free?

Lead sheet: melody + chords



⇒ It is simply a low rhythmicity multi-track sample



⇒ **Lead sheet orchestration** = re-orchestration with higher rhythmicity
(and higher polyphonicity)

oboe-cello-harp-trio

Baseline models

- FIGARO²: Multi-track music style transfer
- AccoMontage-band³: Lead sheet band arrangement

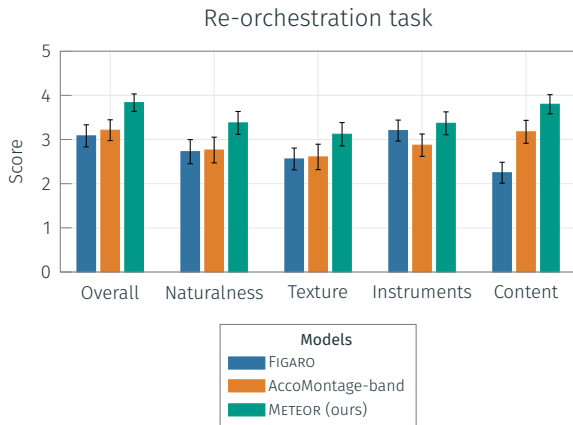
Objective (fidelity + controllability metrics) and **subjective** evaluation.

Criteria: Overall musicality ; Naturalness ; Texture fidelity ; Convincing use of instruments ; Content coherency.

²Rütte & al., *FIGARO: Generating Symbolic Music with Fine-Grained Artistic Control*, 2023

³Zhao & al., *Structured multi-track accompaniment arrangement via style prior modelling*, 2024

Subjective evaluation



Piano solo → orchestra

Same texture

(Rachmaninoff - Prelude in G minor)

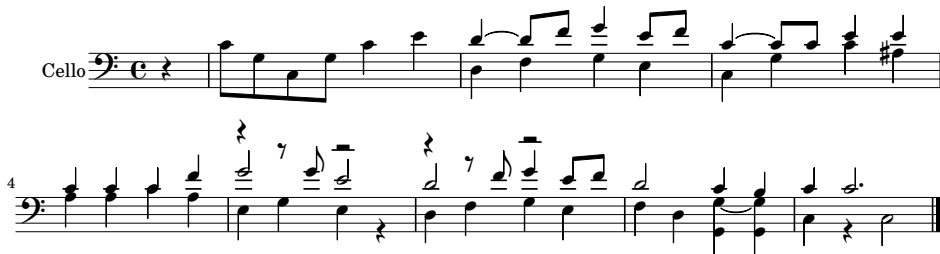
reference

meteor

- **Content coherency:** FIGARO does not keep the melody figaro
- **Texture fidelity:** AccoMontage relies on a lead sheet transcription accomontage

Towards symbolic music generation for *human use*?

piano-trio-reorchestration ⁴



⁴Reference: lead sheet of Auld Lang Syne (*Ce n'est qu'un au revoir*).

Towards symbolic music generation for *human use*?

piano-trio-reorchestration ⁴

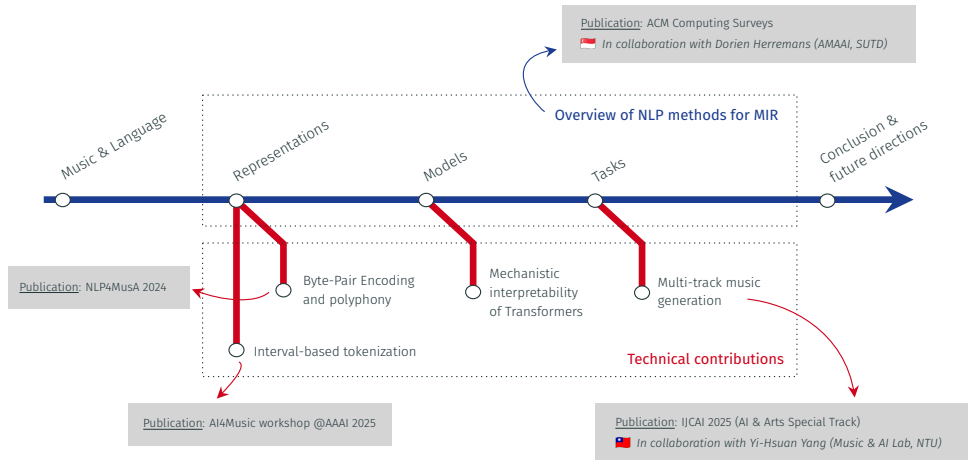
The image shows a musical score for two instruments: Cello and Bass. The Cello part is on a single staff in bass clef with a common time signature (C). The Bass part is on a single staff in bass clef with a common time signature (C). The score is divided into two systems. The first system has a measure where the Cello plays a series of eighth notes, and the Bass plays a series of eighth notes. The second system has a measure where the Cello plays a series of eighth notes, and the Bass plays a series of eighth notes. The score is annotated with several colored boxes: a large orange box highlights a section of the Cello staff, a pink box highlights a section of the Bass staff, and a small orange box highlights a section of the Bass staff. The score is also annotated with a '4' in the top right corner.

- Enharmonics, voice separation,...
 - Awkward fingerings, articulations, breathings,...
- How to ensure **humanly understandable** and **playable** generated music?

⁴Reference: lead sheet of Auld Lang Syne (*Ce n'est qu'un au revoir*).

Future directions & conclusion

Contributions, collaborations & publications



Other publications: Computational musicology (DLfM 2022) ; Music corpora (TISMIR 2025) ; AI for general public (Culture & Recherche 2024).

Musical alphabet vs. text alphabet

- Is the Latin alphabet really the closest to musical alphabet?
- Can one still rely on MIDI to make *playable* music generation?

Data availability: text vs. symbolic music data

- Text data is released much more quickly than symbolic music
- Are larger models really the solution?

Practices in the NLP field for MIR

- **Model explainability:** understanding the behavior of a model is essential, particularly in creative contexts.
 - *In NLP:* model-agnostic tools ; mechanistic interpretability
 - ISMIR 2025 Tutorial “Explainable AI for MIR”
- **Benchmarks:** how to compare models not developed for the same tasks?
 - *In NLP:* common metrics (BLEU for translation, ...) ; standardized benchmarks (GLUE, ...)
 - MIREX competition ; SMC Benchmark [Wang & al. ISMIR 2025]

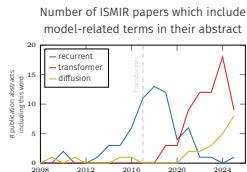
Towards further models for MIR

Practices in the NLP field for MIR

Towards further models for MIR

- Transformers are still used, but have become “standard”
- Adapting models from other fields
 - e.g. diffusion models from image processing

↪ What about a model tailored for music, beyond “simple” **adaptations** of other models?



How can one structure a MIR project?

- **Task:** how precisely is it musically defined?
- **Representation:** what motivates a particular choice?
- **Model:** why is it suitable for the chosen task and representation?

NLP as a *toolbox* for MIR

- NLP tools can be efficient to process symbolic music.
- Applying NLP tools for symbolic music should be first driven by *musical* questions.

How can one structure a MIR project?

- **Task:** how precisely is it musically defined?
- **Representation:** what motivates a particular choice?
- **Model:** why is it suitable for the chosen task and representation?

NLP as a *toolbox* for MIR

- NLP tools can be efficient to process symbolic music.
- Applying NLP tools for symbolic music should be first driven by *musical* questions.

Thanks for listening!

Modeling Symbolic Music with Natural Language Processing Approaches

